Question Answering System Using Syntactic Information

Masaki Murata   Masao Utiyama   Hitoshi Isahara
Intelligent Processing Section, Kansai Advanced Research Center,
Communications Research Laboratory
Ministry of Posts and Telecommunications
588-2, Iwaoka, Nishi-ku, Kobe, 651-2492, Japan
TEL:+81-78-969-2181   FAX:+81-78-969-2189
http://www-karc.crl.go.jp/ips/murata
{murata,mutiyama,isahara}@crl.go.jp

Abstract

Question answering task is now being done in TREC8 using English documents. We examined question answering task in Japanese sentences. Our method selects the answer by matching the question sentence with knowledge-based data written in natural language. We use syntactic information to obtain highly accurate answers.

1 Introduction

Question answering task has been done in TREC8 using English documents [1]. Here, we examine question answering task in Japanese sentences[1,2]. Our approach is to use syntactic information.

2 Question Answering System

2.1 Outline

1. The system detects keywords in question sentences, and then detects sentences in which the sum of the keywords' IDF values is high[3].

2. The question sentences and the detected sentences are parsed by the Japanese syntactic analyzer [8]. (This allows us to obtain the dependency structures.)

3. The answer is selected by matching a question sentence and the detected sentences using syntactic information. How this is done is described in the next section.

[1] With respect to Japanese sentences, domain-dependent work such as that on dialogue systems and help systems has been done [2] [3], but little work has been done on detecting the answer from natural-language databases as in question answering task. However, much has been done on English sentences, such as work on detecting sentences in written answers [4] to work on detecting answers themselves [5].

[2] This paper outlines one part of a question answering system that we have been developing for a long time [6] [7].

[3] In this paper, the system detects one sentence by one sentence. However, it would be better to detect a series of sentences and detect the answer from a series of sentences. If a series of sentences is used to detect the answer, context information can be used.

2.2 Matching a Question and Detected Sentences Using Syntactic Information

We use the syntactic information when matching a question sentence and the detected sentences.

The score of a detected sentence s is as follows.

$$Score(s) = B1(s) + \alpha * B2(s) - \beta * DNUM(s) \tag{1}$$

$$B1(s) = \sum_{\text{all bunsetsus b in the question sentence}} BNST1(b) \tag{2}$$

$$B2(s) = \sum_{\substack{\text{all pairs of two bunsetsus (b1, b2) in} \\ \text{the question sentence, where b1 de-} \\ \text{pends on b2 (i.e. b2 is the head of b1.)}}} BNST2(b1, b2) \tag{3}$$

Each of the bunsetsus in the question sentence can be paired with one of the bunsetsus in the sentence s in order to maximize the value of $Score(s)$. (A *bunsetsu* in Japanese corresponds to a phrasal unit such as a noun phrase or a prepositional phrase in English.) $BNST1(b)$ is the similarity between the bunsetsu b in the question sentence and the bunsetsu in the detected sentence s paired with the bunsetsu b. $BNST2(b)$ is the similarity between the set of the two bunsetsus, ($b1$, $b2$), and the set of the two bunsetsus in the detected sentence s paired with $b1$ and $b2$. $DNUM(s)$ is the number of the bunsetsus of the sentence s. α and β are constants, and are set by experiment. (Although we use only monomial and binomial syntactic information in Eq. 1, we can also use trinomial or polynomial syntactic information.)

We calculate the similarity between two words by using the EDR dictionaries [9][4]. In the case of the bunsetsu containing an interrogative pronoun, the similarity is calculated according to the situations. For example, when a bunsetsu in the question sentence is “where” and the paired bunsetsu in the sentence s has the meaning of location[5], the similarity between them is set to high.

Our system performs the above matching process and selects the answer from the sentence having the highest score. The answer is selected by considering a bunsetsu paired with a bunsetsu containing an interrogative pronoun as the desired answer.

In general, the answer of the question sentence can be obtained by matching the question sentence and the database sentences. In the case of YES-NO questions, the system has only to match the question sentence and the database sentence, and outputs YES if matched (or NO otherwise). In the case of fill-in-the-blanks-type questions[6], the system has only to consider

[4] The similarity between words can be handled by using thesauri. But the similarity between long expressions such as clauses is difficult to handle. To solve this problem, we have already considered the method of using rewriting rules [10]. This method will be described in later papers.

[5] Specifying bunsetsus whose meanings are locations is done by using thesauri such as the EDR dictionaries.

[6] The process of solving fill-in-the-blanks-type questions can be considered as a case of ellipsis resolution if the blanks are considered as ellipses. We have already discussed how corpora can be used in ellipsis resolution [11]. So we should be able to use corpora to fill blanks in the fill-in-the-blanks-type questions.

the element of the database sentence, paired with an interrogative pronoun such as "What" as the desired answer. Our approach here is an implementation of this idea using syntactic information.

3 Example

This section shows three examples of when our system obtained correct answers.

We used as question sentences the English-to-Japanese translations of sample sentences in TREC8. We used as database sentences the Daijirin Japanese word dictionary and the Mainichi Japanese newspaper (1991-1998). When we use the Daijirin dictionary, we added the strings "entry word + *wa* (topic-marking functional word)" to the beginning of each sentence.

First, we inputted the following Japanese sentence into our system.

uganda	*no*	*shuto*	*wa*	*doko*	*desu*	*ka.*
(Uganda)	(of)	(capital)	topic	(where)	(be)	(?)

(What is the capital of Uganda?)

As a result of calculating the score of Eq. 1, the following sentence in the Daijirin dictionary had the highest score and "Kampala" was correctly selected.

kanpara	*wa*	*uganda*	*kyouwakoku*	*no*	*shuto*	*desu.*
(Kampala)	topic	(Uganda)	(republic)	(of)	(capital)	(be)

(Kampala is the capital of the Uganda republic.)

The score is calculated in the following.

$$\begin{aligned} Score &= 9.7(\text{Matching between ``Uganda'' and ``Uganda republic''}) \\ &+ 5.9(\text{Matching between ``capital'' and ``capital''}) \\ &+ 1.6(\text{Matching between ``capital of Uganda'' and ``capital of Uganda republic''}) \\ &\quad ... \\ &= 17.2 \end{aligned} \tag{4}$$

Next, we inputted the following Japanese sentence into our system.

magunakaruta	*ga*	*tyouin*	*sareta*	*no-wa*	*nan*	*nen*	*desu*	*ka.*
(Magna Carta)	subject	(sign)	passive	topic	(what)	(year)	(be)	(?)

(What year was the Magna Carta signed?)

As a result of calculating the score of Eq. 1, the following sentence in the Daijirin dictionary had the highest score and "1215" was correctly selected.

magunakaruta	*wa*	*1215*	*nen*	*igirisu*	*no*	*houken*	*shokou*	*ga*
(Magna Carta)	topic	(1215)	(year)	(England)	(of)	(feudal)	(lords)	subject

kokuou	*jon*	*ni*	*semari,*	*ouken*	*no*	*seigen*	*to*
(king)	(John)	object	(press)	(royal authority)	(of)	(limitation)	(and)

shokou	*no*	*kenri*	*wo*	*kakunin*	*saseta*	*bunsho.*
(lords)	(of)	(right)	object	(confirm)	causative	(document).

(Magna Carta is the document in which feudal lords of England made King John confirm the limitation of the royal authority and their rights in 1215.)

The score is calculated in the following.

$$\begin{aligned} Score &= 32.0(\text{Matching between } \textit{nan nen} \text{ “what year” and } \textit{1215 nen} \text{ “1215 year”}) \\ &+ 14.6(\text{Matching between “Magna Carta” and “Magna Carta”}) \\ &\quad ... \\ &= 48.1 \end{aligned} \tag{5}$$

Finally, we inputted the following Japanese sentence into our system.

paakinson	*byou*	*wa*	*nou*	*no*	*dono*	*bubun*	*ni-aru*	*saibou*	*no*
(Parkinson)	(disease)	topic	(brain)	(of)	(what)	(area)	(in)	(cell)	(of)

shi	*ni*	*kankei-shite-imasu*	*ka.*
(demise)	(to)	(be linked)	(?)

(The symptoms of Parkinson's disease are linked to the demise of cells in what area of the brain?)

As a result of calculating the score of Eq. 1, the following sentence in the Mainichi newspaper had the highest score and “substantia nigra” was correctly selected.

paakinson	*byou*	*wa*	*tyuunou*	*no*	*kokushitsu*	*ni-aru*	*meranin*
(Parkinson)	(disease)	topic	(midbrain)	(of)	(substantia nigra)	(in)	(melanin)

saibou	*ga*	*hensei-shi,*	*kokusitsu*	*saibou*	*nai-de*	*tsukurareru*
(cell)	subject	(degenerate)	(substantia nigra)	(cell)	(in)	(be made)

shinkei-dentatsu-busshitsu	*no*	*doupamin*	*ga*	*nakunari*	*hatsubyou-suru,*
(neurotransmitter)	(of)	(dopamine)	subject	(run out)	(be taken ill)

to-sarete-iru.
(be recognized)

(Parkinson's disease is recognized when melanin cells in the substantia nigras of the midbrain degenerate. The neurotransmitter dopamine, which is made in substantia nigra cells, runs out, and Parkinson's disease arises.)

The score was calculated in the following.

$$\begin{aligned} Score &= 10.6(\text{Matching between “Parkinson's disease” and “Parkinson's disease”}) \\ &+ 6.3(\text{Matching between “cell” and “melanin cell”}) \\ &+ 1.5(\text{Matching between “brain” and “midbrain”}) \end{aligned} \tag{6}$$

$$
\begin{aligned}
& \ldots \\
+ \quad & 0.4(\text{Matching between “area of brain” and “substantia nigra of midbrain”}) \\
+ \quad & 0.3(\text{Matching between “in area” and “in substantia nigra”}) \\
& \ldots \\
= \quad & 32.2
\end{aligned}
$$

“cells in interrogative pronoun of brain” and “cells in substantia nigra of midbrain” were matched, and “substantia nigra” was correctly selected[7].

4 Conclusion

We have outlined our question answering system using syntactic information. We intend to run more experiments, to make our system more robust.

We think that the human sentence-reading process involves a matching process between the sentence being read now and data recalled in the brain [12]. Our question answering system matches question sentences and sentences in its database, and may therefore provide some clues to shed light on the human reading process. Future work will involve extending this current work to work on the human reading process.

References

[1] TREC-8 Question Answering Track, http://www.research.att.com/~singhal/qa-track.html, (1999).

[2] Tadahiko Kumamoto and Akira Ito, An analysis of real and simulated dialogues in the same task domain, *Proc. of Pacific Association for Computational Linguistics (PACLING'99)*, (1999).

[3] Wataru Higasa and Sadao Kurohashi, Dialogue helpsystem based on flexible matching of user query with knowledge-base, *Information Processing Society of Japan, WGNL-133*, (1999), (to appear), http://www-lab25.kuee.kyoto-u.ac.jp/member/higasa/research/index.html, (in Japanese).

[4] Thomas S. Morton, Using coreference for question answering, *ACL Workshop, Coreference and Its Applications*, (1999).

[5] Julian Kupiec, MURAX: A robust linguistic approach for question answering using an on-line encyclopedia, *In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1993).

[6] Masaki Murata, Dialogue and natural language understanding, (unpublished manuscripts, October 1994), (in Japanese).

[7] Masaki Murata, Question answering system using large scale data written in natural language, (unpublished manuscripts, April 1999), (in Japanese).

[8] Sadao Kurohashi, *Japanese Dependency/Case Structure Analyzer KNP version 2.0b6*, (Department of Informatics, Kyoto University, 1998), (in Japanese).

[7] It would be good to modify Eq. 1 in order to increase the similarity of bunsetsus around an interrogative pronoun.

[9] EDR (Japan Electronic Dictionary Research Institute, Ltd.), *EDR Electronic Dictionary Technical Guide*.

[10] Noam Chomsky, Three models for the description of language, *IRE Transactions on Information Theory*, Vol. 2, No. 3, (1956), pp. 113–124.

[11] Masaki Murata and Makoto Nagao, Anaphora/ellipsis resolution method using surface expressions and examples, *IEICE-WGNLC97-56*, (1998), (in Japnese).

[12] P.N. Johnson Laird, *Mental models*, (Cambridge Univ. Press, 1983).